HierarX : un outil pour la découverte de hiérarchies dans des espaces hyperboliques à partir de similarités

François Torregrossa*,**, Guillaume Gravier**
Vincent Claveau**, Nihel Kooli*

*Solocal, Rennes 35000
{ftorregrossa,nkooli}@solocal.com,
http://www.solocal.com
**IRISA, Rennes 35000
francois.torregrossa@irisa.fr
guillaume.gravier@irisa.fr
vincent.claveau@irisa.fr
http://www.irisa.fr

Résumé. Cet article introduit l'outil HierarX, permettant de projeter des données dans des espaces hyperboliques : Lorentz ou Poincaré. À partir de similarités entre des mots ou de leur représentation dans des espaces de grandes dimensions, HierarX incorpore des connaissances dans des géométries hyperboliques de petites dimensions. Ces dernières ont la particularité de représenter l'information sous forme de hiérarchies continues. Ce travail présente le fonctionnement de HierarX ainsi que ses principaux cadres d'utilisation.

1 Introduction

Les plongements de mots ou *word embeddings* sont des méthodes bien connues et largement utilisées pour la compréhension du langage naturel. À partir de diverses sources, ces solutions proposent de calculer la représentation continue des mots ou concepts composant le langage. En finalité, des relations géométriques, utiles aux systèmes automatiques, en émergent.

En pratique, les sources de données sont multiples : des phrases brutes ou encore des similarités. Word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) et FastText (Bojanowski et al., 2017) sont par exemple des méthodes travaillant sur des phrases brutes en vue de découvrir des relations paradigmatiques ou syntagmatiques. D'autres méthodes étudient les similarités entre paires de mots, par exemple les travaux de Sun et al. (2015) et plus récemment ceux de Nickel et Kiela (2018). Leur proposition est de reconstruire des données hiérarchiques en utilisant les similarités entre éléments, qui forment une source de données plate et déstructurée. Malgré des résultats prometteurs, les implémentations des deux solutions n'ont pas été rendues publiques. De plus, celles-ci sont particulièrement difficiles à réaliser en pratique, puisqu'elles font intervenir des espaces hyperboliques, plus adaptés à la représentation des hiérarchies de façon continue. HierarX est le logiciel que nous proposons pour combler ce manque.